



## **Feasibility of estimating travel demand using geolocations of social media data**

Downloaded from: <https://research.chalmers.se>, 2023-05-04 22:15 UTC

Citation for the original published paper (version of record):

Liao, Y., Yeh, S., Gil, J. (2022). Feasibility of estimating travel demand using geolocations of social media data. *Transportation*, 49(1): 137-161. <http://dx.doi.org/10.1007/s11116-021-10171-x>

N.B. When citing this work, cite the original published paper.



# Feasibility of estimating travel demand using geolocations of social media data

Yuan Liao<sup>1</sup> · Sonia Yeh<sup>1</sup> · Jorge Gil<sup>2</sup>

Accepted: 11 January 2021  
© The Author(s) 2021

## Abstract

Travel demand estimation, as represented by an origin–destination (OD) matrix, is essential for urban planning and management. Compared to data typically used in travel demand estimation, the key strengths of social media data are that they are low-cost, abundant, available in real-time, and free of geographical partition. However, the data also have significant limitations: population and behavioural biases, and lack of important information such as trip purpose and social demographics. This study systematically explores the feasibility of using geolocations of Twitter data for travel demand estimation by examining the effects of data sparsity, spatial scale, sampling methods, and sample size. We show that Twitter data are suitable for modelling the overall travel demand for an average weekday but not for commuting travel demand, due to the low reliability of identifying home and workplace. Collecting more detailed, long-term individual data from user timelines for a small number of individuals produces more accurate results than short-term data for a much larger population within a region. We developed a novel approach using geotagged tweets as attraction generators as opposed to the commonly adopted trip generators. This significantly increases usable data, resulting in better representation of travel demand. This study demonstrates that Twitter can be a viable option for estimating travel demand, though careful consideration must be given to sampling method, estimation model, and sample size.

**Keywords** Social media data · Travel demand · Origin–destination estimation · Longitudinal data · Lateral data · Gravity model

---

✉ Yuan Liao  
yuan.liao@chalmers.se

<sup>1</sup> Division of Physical Resource Theory, Department of Space, Earth and Environment, Chalmers University of Technology, Gothenburg, Sweden

<sup>2</sup> Division of Urban Design and Planning, Department of Architecture and Civil Engineering, Chalmers University of Technology, Gothenburg, Sweden

## Introduction

Travel demand estimation is essential for urban planning and management of transportation networks. The time series of visits to various locations by individuals are aggregated to study the flows of people between different zones/regions. Based on the spatio-temporal scale of the aggregation, an origin–destination (OD) matrix can be constructed with the origins and destinations of all trips. These OD matrices are particularly important for representing travel demand (Calabrese et al. 2011). Traditionally, the estimation of OD matrices relies on input data from household travel surveys, censuses, and traffic surveys that feature representative populations and detailed information about travel mode and trip purposes. However, data collection frequency, methods, and data availability vary across countries (and across cities within a country), making it difficult to interpret the results. For example, in the UK and the Netherlands the travel surveys are done annually, but that is an exception. Other places do not do them regularly, if at all. Portugal had one travel survey carried out for two metro areas in 2017, but nothing more since. Otherwise, mobility is derived from the census data (carried out every 10 years), but that offers a different resolution since it is not based on travel diaries. On top of these issues, the costs of these surveys are increasing, while the response rates are decreasing over time (Yue et al. 2014), making it hard to keep the travel demand models up to date. Emerging data sources associated with mobile/smart phones are increasingly leveraged to overcome these drawbacks.

In the last decade, the emerging data sources have significantly improved our understanding of travel behaviour (Gonzalez et al. 2008; Song et al. 2010; Barbosa et al. 2018) and have brought new opportunities for travel demand modelling (Anda et al. 2017). Common emerging data sources are call detail records (CDR) (Calabrese et al. 2011), smart card data, GPS-enabled devices, and geotagged social media, e.g., Twitter (Lee et al. 2019; Hasnat and Hasan 2018).

Alongside the development of information and communication technologies (ICT), interest in online social media services, e.g. Twitter, has grown among the transportation research community (Rashidi et al. 2017). A tweet typically contains multiple components that can be useful for transport research, including text, hashtag, location, and timestamp. When users choose to have their location reported when sending out tweets, these are called *geotagged tweets*. Despite geotagged tweets accounting for a small proportion (1–3%) of all tweets (Morstatter et al. 2013), these check-ins provide precise location information and have increasingly been used for estimating mobility and travel demand either at the global (e.g. Hawelka et al. 2014) or regional level (e.g. Yang et al. 2015).

In the estimation of travel demand, two forms of data are often used: longitudinal and lateral. A *longitudinal* data set is characterised by long-term (more than 24 h) and continuous observations focusing on a group of participants. A *lateral* data set is often collected based on a particular area, such as a city or a country, during a short to medium time period, and it usually covers a larger population. Thus, the data offer either broader or longer coverage, but rarely both.

*Geotagged tweets* can be obtained in three ways: (1) Purchase the complete set of public tweets from Twitter Firehose (Twitter 2019c); (2) Access the Streaming API to get a maximum of 1% of the public tweets (Twitter 2019a); (3) Access the user timeline by user name/ID to get a maximum of 3200 historical tweets that are set by the user as publicly accessible (Twitter 2019b). Different collection channels of geotagged tweets

correspond to different data forms. Sampling methods (1) and (2) collect geotagged tweets generated within a specified region, while sampling method (3) collects data from user timelines without any spatial boundaries.

Geotagged tweets collected from Twitter Firehose and Streaming API are often limited to a geographical bounding box yielding a lateral data set. It covers a large number of Twitter users but takes time to accumulate enough samples for each individual, and movements outside or across the bounding box are not captured (Liao et al. 2019). Alternatively, by accessing User Timeline API, all publicly available historical tweets by a specific user can be collected to form a longitudinal record of individual trajectories without any geographical boundaries. Longitudinal geotagged tweets are collected without being constrained to a specific area, but typically with a smaller number of individuals, albeit a much larger overall sample size (one to two orders of magnitude more samples per user).

Most studies use geotagged tweets in the lateral form, focusing on a specified area in line with the spatial scale of policy-making and urban planning. For example, one study modifies a classic movement model by integrating locations posted on Foursquare (which Twitter integrates) for origin–destination estimation in Austin, Texas (Jin et al. 2014). Longitudinal data can also be scaled up to large numbers of Twitter users to study the OD flows between global cities (Lenormand et al. 2015).

One recent literature review shows that experts are optimistic about the usefulness of such data sources for modelling travel behaviour (Rashidi et al. 2017). Compared with the other data sources, geotagged tweets have several strengths: long collection duration, large number of studied individuals, large spatial coverage, ease of access, low cost, and accurate location information. The low cost of retrieving geotagged tweets makes them especially appealing compared to other data sources (Rashidi et al. 2017). The data source is free to access, and it provides precise location information with a spatial resolution of around 10 m compared with 100–200 m for call detail records (CDR) (Jurdak et al. 2015). Moreover, it is relatively scale free, i.e. analyses can be done with any desired time frame and spatial boundaries based on the research question at hand (Liao and Yeh 2018).

Despite the wide applications, rigorous cross-validation of the use of emerging data sources, such as geotagged social media data, to approximate the travel demand, and their robustness across spatial and temporal scales is still lacking. The main criticism of Twitter data pertains to two aspects: a biased population representation, and low and irregular sampling. Geotagged tweets can capture movements over multiple years and include overseas visits, but the data are “sparse”, thus the picture of actual movements is incomplete (Liao et al. 2019). There have been studies comparing multiple data sources to identify/adjust the biases (e.g. Wesolowski et al. 2013; Tasse et al. 2017) and to validate against “ground truth” (e.g. Lee et al. 2019). It is worth noting, however, that the “ground truth” is also an incomplete picture of reality, as it is, at best, based on the knowledge from well-recognised but limited data collection and established modelling techniques.

This study attempts to comprehensively examine the validity of using geotagged Twitter data for travel demand estimation by comparing Twitter data sets with established data sources. We first compare the empirical trip records with respect to the commuting travel demand and the overall travel demand for an average weekday. We then create gravity models based on Twitter data to estimate the overall travel demand at both the national (long-distance travel above 100 km) and city level. Finally, we compare Twitter-based OD matrices and trip distance distributions with those from the other established sources using spatially weighted structural similarity index and Kullback–Leibler divergence, respectively.

The main contributions of this study lie in the quantification of the feasibility of using geolocations of Twitter data for estimating commuting demand and the overall travel demand, given different sample sizes, sampling methods of Twitter data, and spatial scales. In addition, we develop a novel approach using geotagged tweets as attraction generators as opposed to the commonly adopted trip generators. This significantly increases usable data, resulting in better representation of travel demand and the promise for using Twitter data at a finer spatiotemporal resolution.

The remainder of this paper is organised as follows. “[Related work](#)” section reviews work related to travel demand estimates using social media data and outlines the objectives of the present study. “[Data description](#)” section describes the data, and “[Methodology](#)” section describes the methods used. The results are presented in “[Results](#)” section, and “[Discussion](#)” section discusses the findings. “[Conclusion](#)” section concludes and identifies future research needs.

## Related work

### Modelling travel demand

For travel demand estimation, one needs to first extract activities and trips where Twitter data have proven useful for both conventional four-step modelling and activity-based modelling by providing inferred activities and trips. There has been increased interest in developing methods to infer this information using social media check-in data, such as Twitter data. One recent study has demonstrated that Twitter data can be integrated with an household travel survey to improve the quality of OD matrices (Cheng et al. 2020). Constructing activity-based models requires trip purpose, departure time, and socioeconomic attributes of travellers, among other attributes. The content of geotagged tweets is often used with text mining to extract those attributes, e.g., the activity purposes such as work and leisure and the socio-economic profile of Twitter users (Hasan and Ukkusuri 2014; Abbasi et al. 2015; Maghrebi et al. 2015).

The methodology of four-step travel demand modelling (McNally 2007) consists of *trip generation* and *trip distribution* as the first two steps. It starts from the definition of a *trip*, which is the connection between two consecutive stays generated by the same individual. This individual refers to a phone user when using CDR data (Calabrese et al. 2011), or a survey participant from a one-day travel diary. When it comes to geotagged social media data, a trip is generally defined in the literature as the connection between two consecutive geotagged tweets generated by the same Twitter user. However, due to the sparsity and incomplete trajectory of geotagged tweets, the time interval between two consecutive geotagged tweets can be extremely long (from a few hours to several weeks/months), while the air distance can be close to zero. Therefore, in this context, “*displacement*” is a more appropriate term than the traditional sense of the trip. Despite a displacement in geotagged tweets being different from a record in a travel diary, existing literature often uses these two terms interchangeably.

### Trip generation

Trip generation involves the estimation of the number of trips produced by and attracted to each zone, either using empirical data directly, or modelled results based on zonal demographics and land use information.

Social media data such as displacements in Twitter data need to be processed to become trips. Gao et al. (2014); Kheiri et al. (2015) and Lee et al. (2019) propose *displacement conversion* where they filter out those displacements with time intervals longer than a selected time threshold, e.g. 4 h, 12 h, or 24 h. However, this time threshold is arbitrary and the choice results in a massive reduction of available data.

Instead of geotagged displacements, one can model destination choices to estimate zonal attractiveness. Hasnat et al. (2019) applied Twitter data together with census tract data for modelling travellers' destination choice behaviour, which suggests that Twitter data can be utilised effectively for modelling destination choices that reflect the attractions of zones. Molloy and Moeckel (2017) develop a long-distance destination choice model using Foursquare check-ins whose results suggest that check-ins from social media platforms can improve destination choice models, particularly for leisure travel.

### Trip distribution

Trips are further aggregated to OD zones depending on the spatial scale. The step of trip distribution assigns trips produced by each zone to each of the other zones where these trips are attracted to Anda et al. (2017). There are many models to assign the number of trips between each pair of OD zones. In a study by Yang et al. (2015) of the Chicago metropolitan region, daily check-ins from Foursquare are used to estimate the productions and attractions in each traffic analysis zone as inputs to gravity models for estimating trip distribution. By further calibrating against the OD matrix from other data sources such as CDRs, they demonstrate how to use gravity models with check-in data to estimate the OD matrix. Kheiri et al. (2015) use the radiation model, rank-based model, and population-weighted opportunities model to distribute the trips generated with Foursquare check-ins to estimate the OD matrix.

### Commuting travel demand estimation

Estimating the OD matrix according to trip purpose points toward more specific applications. Commuting flows account for a large share of total trips, therefore they attract more attention. For example, Zagatti et al. (2018) use CDRs to estimate an OD matrix of commuting flows. For social media data, some data sources have trip purposes (activity types), such as Foursquare, while Twitter data do not directly provide this information. With a small share of check-ins at home/workplace from Foursquare when compared with the actual daily mobility, Yang et al. (2015) focus on non-commuting trips. To construct OD matrices of commuting flows with geotagged tweets or CDRs, one needs to detect home/workplace when the trip purpose is not explicitly given. Schneider et al. (2013) assume that the most visited location during weekends and 7 pm–8 am on weekdays is the home location and the second most visited location during 8 am–8 pm on weekdays is identified as one's workplace. Combining such temporal rules and visiting frequency, this method has been widely used to identify the home/workplace through social media data (Wang et al. 2018; Osorio-Arjona and García-Palomares 2019), sometimes together with land-use information (Osorio-Arjona and García-Palomares 2019).

Efforts that infer the home/workplace from geotagged tweets must consider the behavioural bias of people geotagging consciously and intentionally in uncommon places to communicate and show where they have been (Tasse et al. 2017). Home and workplace are at the opposite extreme, i.e., they are the most common places that people visit on a

daily basis. A preliminary comparison between Twitter data and the national travel survey suggests that the low probability of reporting home and workplace implies that further scrutiny of the validity of estimating commuting-OD matrices based on geotagged tweets is required.

## Validation against other data sources

Researchers have devoted efforts to validating geotagged tweets with other data sources. A study focusing on the U.S. found that densely populated regions and males were over-represented among Twitter users (Mislove et al. 2011). In addition, there are two possible types of behavioural distortion for Twitter users who geotag: only tweeting at specified locations or times, and geotagging only certain or all of the tweets.

When cross-validating against data with higher temporal resolution such as CDR (Lenormand et al. 2014), good agreement is generally found regarding, for instance, trip distance distribution. When validating geotagged tweets against travel surveys, studies show that geotagged social media data capture the displacement distribution, length, duration, and start time of trips reasonably well for the purpose of inferring individual travel behaviour (Zhang et al. 2017; Liao et al. 2019). Validations using CDR need careful interpretation, as CDR and geotagged tweets are both passive data collection methods that share some similar shortcomings.

Good agreement on fundamental indicators of individual travel behaviour does not necessarily guarantee a good proxy for the travel demand at the population level. Some studies comparing geotagged tweets with traffic data (Ribeiro et al. 2014) and travel-demand data (Lee et al. 2015, 2019; Yang et al. 2015) have generally achieved good results. However, as pointed out recently by Lee et al. (2019), the sparsity of geotagged tweets leads to sparse OD matrices and therefore cannot replace other travel demand forecasting methods for state-wide travel models.

## Study objectives

The work comparing geotagged tweets with other data sources for travel demand estimation still lacks systematic rigour in at least four areas: (1) *Commuting travel demand*. The basic temporal technique to identify home or workplace has been widely applied for deriving commuting trips. Our preliminary results from previous analyses suggest that identifying home and workplace locations through geotagged tweets gives mixed results and the reliability of the method requires further scrutiny; (2) *Spatial scale*. Most studies look at pre-selected regions without exploring the effects of spatial scales on travel demand estimation, whereas we hypothesise that the feasibility of using Twitter data for travel demand estimation can depend on the scale; (3) *Sampling methods*. The existing literature is not clear on how different sampling methods (region-based vs. user-based) affect the validity of using geotagged tweets to estimate travel demand; (4) *Sample size*. It remains unclear how the sparsity of Twitter data affects the validity of using it for travel demand estimation.

To fill these gaps in the literature, we systematically examine the validity of using geotagged tweets collected within a specified region, and from user timelines, to approximate the OD matrix at different spatial scales. We compare these Twitter-based OD matrices with the Swedish national travel survey and output from Swedish Transport Administration (Trafikverket) traffic models. Specifically, we attempt to answer the following questions:



- Are Twitter data a feasible source for representing commuting travel demand?
- Can geolocations of Twitter data be used to create models for travel demand estimation?
- How do spatial scale, sampling method, and sample size of Twitter data affect its representativeness for travel demand?

## Data description

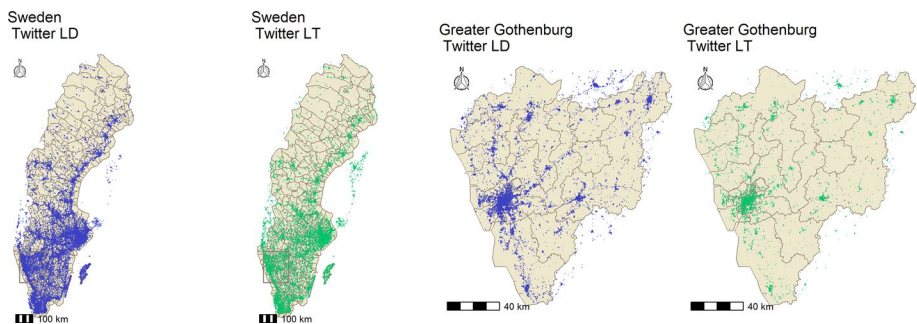
This study focuses on Sweden as a whole and on Greater Gothenburg, located in western Sweden. Sweden is a European country with a population of 10.2 million in 2019 and the GDP per capita was 54.6 kUSD in 2018 (Statistics Sweden). Gothenburg is its second largest city for which Greater Gothenburg covers its metropolitan area with a population of around 1 million.

Specifically, four datasets have been used in this study. Two Twitter datasets collected using different sampling methods: lateral geotagged tweets (Twitter LT), and longitudinal geotagged tweets (Twitter LD). And two datasets with which the Twitter data are compared: the Swedish National Travel Survey; and OD matrices from the Sampers model, a traffic simulation model with the travel demand module embedded, developed by the Swedish Transport Administration. The traffic zones used by Sampers are illustrated in Fig. 1 for two spatial scales: Greater Gothenburg (city level) and Sweden (national level). Detailed descriptions of each dataset are presented in this section.

## Twitter data

### Lateral geotagged tweets (Twitter LT)

We purchased data from Gnip, a Twitter subsidiary, during a 6-month period (20 December 2015–20 June 2016) within the geographical bounding box of Sweden (Jeuken 2017; Liao et al. 2019). Gnip sells complete historical tweets in bulk and provides access to the Firehose API.



**Fig. 1** Geotagged tweets in traffic zones from Twitter LD (blue) and LT (green) for Sweden (left) and Greater Gothenburg (right). (Color figure online)



## Longitudinal geotagged tweets (Twitter LD)

We identify 7773 top geotag users from Twitter LT who geotagged their tweets most frequently during that 6-month period. We extract those top users' historical tweets using Twitter User Timeline API, without applying a spatial boundary limit. This method has a maximum number of tweets that can be collected from a specified user, producing varied time spans and varied tweet numbers, as not all users reached the 3200-tweet maximum.

## Preprocessing and statistics of Twitter data

All the geotagged tweets are preprocessed to reduce potential artefacts causing biases in travel demand estimation. First, we only keep tweets that were generated from mobile devices. Moreover, those users who only had geotagged tweets of a single place are removed due to being bot accounts, e.g., for job posting or weather updates (Ek and Wennerberg 2020). Next, Twitter users can cross-post geotagged tweets from other social media platforms, yielding a place's location being posted instead of the tweet's precise geolocation, for example, the centre of Sweden or the centre of Gothenburg. These geotagged tweets without precise GPS coordinates are also removed. Finally, two filters are implemented for Twitter LD only. The top geotag Twitter users who have less than 50 geotagged tweets in total are removed. Considering the long time span of a given Twitter user's Twitter timeline, he/she might have migrated from one country to another. To avoid confusion, we only keep the latest time period of the geotagged tweets where a Twitter user is assumed to live in Sweden. For the national level, all the Twitter LT and LD are used while for the city level, only these geotagged tweets within the boundary of Greater Gothenburg are used.

We derive the home and workplace locations from Twitter LD given the larger numbers of geotagged tweets per user. The home location is identified as the most-visited location on weekends and between 7pm and 8am on weekdays, whereas the most visited non-home location between 8am and 8pm on weekdays is identified as the user's workplace (Schneider et al. 2013; Wang et al. 2018; Osorio-Arjona and García-Palomares 2019).

Following the practice in the literature to account for the fact that Twitter users are not representative of the overall population, we give weights for individual Twitter users in Twitter LD. The weight is the ratio of Twitter users to the true population in the municipality (Wang et al. 2018). The trips of the Twitter users in Twitter LD are aggregated and multiplied with their individual weight to derive a population-level travel demand estimation. The Twitter users' distributions are found to correlate with the census (Kendall's tau = 0.65,  $p < 0.001$ ). However, top Twitter users tend to be over-represented in big cities especially the top three cities in Sweden: Stockholm (Twitter = 18% vs. Census = 9.4%), Gothenburg (6.9% vs. 5.6%), and Malmö (4.5% vs. 3.3%).

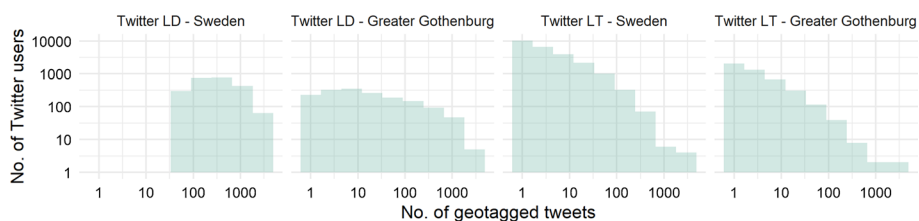
The basic statistics of Twitter LT and LD are summarised in Table 1. Compared with Twitter LT, Twitter LD collected from user timelines without using any spatial bounding box covers a longer time span, contains a larger volume of geotagged tweets and a higher number of geotagged tweets per user and in total, but covers a smaller population than Twitter LT. The distribution of the number of total geotagged tweets per user is shown in Fig. 2.

**Table 1** Statistics of Twitter data used in this study

	Twitter LD	Twitter LT
Time span	2010-09–2019-03	2016-06–2016-12
No. of Twitter users	2311 (1637)	24,442 (4521)
No. of geotagged tweets (million)	11.5 (1.4)	3.4 (0.5)
Days covered per <sup>a</sup>	191 (6)	2 (1)
No. of geotagged tweets per covered day per <sup>a</sup>	1.4 (1.3)	1.0 (1.0)

The numbers in brackets are for Greater Gothenburg

<sup>a</sup>The median value of all Twitter users

**Fig. 2** Distribution of the total number of geotagged tweets of Twitter users by spatial scale and dataset

## Swedish national travel survey (Survey)

The survey data come from the Swedish National Travel Survey (one-day travel diary) for the years of 2011 to 2016 (Official Statistics of Sweden 2016). It consists of a total of 171,553 trips from 38,258 participants covering 2189 record days, with detailed information on individual trip's origin and destination, distance, travel time, and participant's home/workplace. The spatial accuracy is the municipality level.

## Model-based travel demand estimations (Sampers)

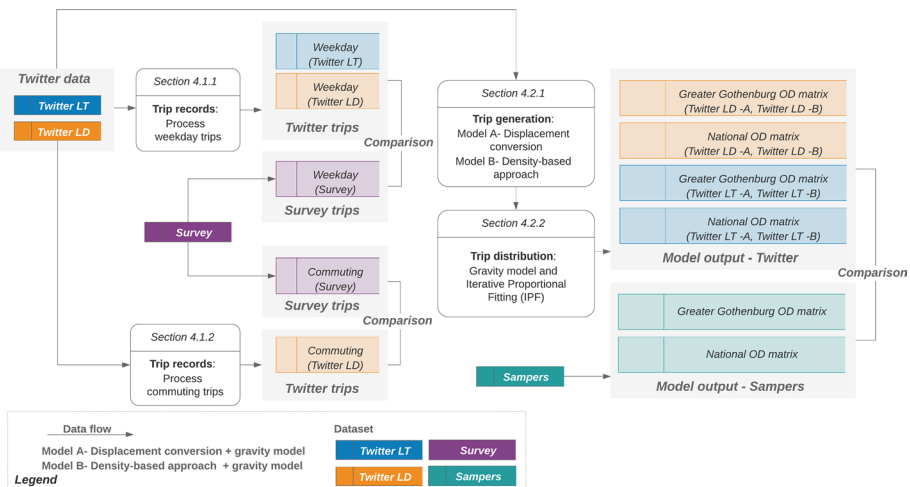
The Swedish Transport Administration uses the Sampers model to calculate changes in traffic volumes under different scenarios. Both the city level and the national level have their own traffic analysis zones that follow the census boundaries and homogeneous socioeconomic characteristics. These spatial zones are used for creating OD matrices with Twitter data so that we can compare Twitter with Sampers' model output.

Sampers calculates travel demand based on studies of travel habits derived from travel surveys, looking at where, how and how often people want to travel, which forms the OD matrices. The model output represents the total travel demand for an average weekday. We used the latest OD matrices (2014) from Sampers for Greater Gothenburg and the entire Sweden. At the national level, we focus on the long-distance trips of Sampers model ( $\geq 100$  km).

## Methodology

In order to examine the feasibility of using Twitter data for travel demand estimation, we use an analytic framework to compare Twitter with the other established data sources, as shown in Fig. 3. In practice, transport planners collect empirical trip data from a small sample of the population and create a model to simulate the travel demand of the overall population for further application, such as traffic flows modelling. Therefore, we divide the comparison into two focuses: empirical trip records (“[Trip records](#)” section) and model output (“[Travel demand model construction](#)” section).

We first compare the empirical trip records obtained from Twitter with those from travel survey data with respect to the overall travel demand for an average weekday (“[Processing weekday trips](#)” section) and commuting travel demand (“[Processing commuting trips](#)” section). In this part of the validation, we also examine the stability of the similarity between Twitter and the travel survey over time. After the analysis of the empirical trips, we create the gravity models, based on Twitter data collected with two sampling methods, to simulate the overall travel demand at both the national (long-distance travel above 100 km) and city level. We use two methods for the step of trip generation (“[Trip generation](#)” section) followed by the gravity model for the trip distribution (“[Trip distribution](#)” section); they are trips converted from displacements by adding a time threshold (*Model A*) and the density-based approach proposed in this study (*Model B*). Model B is proposed as an alternative to Model A to solve the sparsity issue of Twitter data. Finally, we evaluate the results (“[Evaluation of Twitter OD matrices](#)” section) by comparing the Twitter-based trips and model outcomes with those from the national travel survey (Survey) and the Sampers model. The techniques used for the comparison include visualisation, similarity measure (“[Spatially weighted structural similarity index](#)” section), and trip distance distribution.



**Fig. 3** The methodology used for examining the feasibility of Twitter data for travel demand modelling. Twitter LD is described in “[Longitudinal geotagged tweets \(Twitter LD\)](#)” section. Twitter LT is described in “[Lateral geotagged tweets \(Twitter LT\)](#)” section. Survey is described in “[Swedish National Travel Survey \(Survey\)](#)” section. Sampers is described in “[Model-based travel demand estimations \(Sampers\)](#)” section

## Trip records

### Processing weekday trips

We define geotagged displacements by connecting every two consecutive geotagged tweets generated by the same user. To convert these displacements into trips, a time threshold can be used to filter out those displacements that have a time interval longer than a predefined threshold (Gao et al. 2014; Kheiri et al. 2015; Lee et al. 2019). We select 270 min, i.e., the 99th percentile of travel time between municipalities from Survey, as the time threshold for the national-level trip generation. For the city-level trip generation, the time threshold of 140 min is selected which is the 99th percentile of travel time within the corresponding county where most parts of Greater Gothenburg are located.

Survey contains complete sets of trip records at the municipality level. By directly aggregating the weekday records, we get the OD matrix of the overall trip records for an average weekday. The low cost of collecting Twitter data makes it easier to keep them updated over time. However, their actual use also depends on the stability of the similarity between Twitter trips and Survey trips over time. Instead of aggregating the records available, we look into the similarity of OD matrices from 2011 to 2016 at the national level by aggregating the records yearly.

### Processing commuting trips

To construct commuting flows with Twitter LD, we define trips connecting home (origin) and workplace (destination), and aggregate those trips at the municipality level. This gives the national commuting OD matrix based on Twitter LD.

To compare a Twitter-based commuting OD matrix we need to construct an equivalent Survey-based OD matrix. Survey has the home and workplace of each participant at the municipality level, and each participant is assigned an individual weight standing for the representativeness of his/her socio-demographic profile in the overall Swedish population, regarding the time period of participating the survey, region, age, and gender. Specifically, the weight is designed as the ratio between the population and the survey respondent in the respective stratum. By linking home and workplace as a commuting trip for a given individual, multiplied by his/her individual weight, we aggregate all the commuting trips and construct the national commuting OD matrix.

## Travel demand model construction

This section introduces the method of taking empirical trips to create modelled output of travel demand (OD matrix). The method consists of two steps, trip generation (“[Trip generation](#)” section) and trip distribution (“[Trip distribution](#)” section).

### Trip generation

- Displacement conversion (Model A)

By aggregating all the trips converted from displacements (see “[Processing weekday trips](#)” section), we get the overall OD matrices for an average weekday at the national level and the city level.

Based on the OD matrix that is aggregated directly from the trip records, the productions,  $P_i$  for a given origin zone  $i$ , are expressed as the summation of  $f_{ij}$ , i.e., the number of trips between the origin zone  $i$  and the destination zone  $j$ , over the destination zone  $j$  ( $P_i = \sum_{j=1}^N f_{ij}$ ,  $i = 1, 2, \dots, N$ ), where  $N$  is the total number of zones. Similarly, trip attractions are expressed as  $A_j^0 = \sum_{i=1}^N f_{ij}$ ,  $j = 1, 2, \dots, N$ .

- Density-based approach (Model B)

Given the sparsity of Twitter data, the reduction of available data due to adding a time threshold further limits the use of geotagged tweets to represent the travel demand at higher granularity. Therefore, we propose an alternative way utilising all the available geotagged tweets and the census data.

The population of each zone (Pop) represents its productions, i.e.,  $P_i = \text{Pop}_i$ ,  $i = 1, 2, \dots, N$ . And the attractions are represented by the number of geotagged tweets in each zone ( $f_j$ ), i.e.,  $A_j^0 = f_j$ ,  $j = 1, 2, \dots, N$ .

For both methods, the trip productions and attractions are balanced so that they sum up to the same number.

### Trip distribution

The gravity model was first proposed for the estimation of an OD matrix in the 1940s (Zipf 1946) and later became one of the most applied methods for the trip distribution (Yang et al. 2015). Given that this study focuses on the data source, we select the following form of gravity model to avoid model complexity:

$$T_{ij} = P_i \frac{A_j F_{ij}}{\sum_{j=1}^N A_j F_{ij}} \quad (1)$$

where  $T_{ij}$  is the number of trips between the origin zone  $i$  and the destination zone  $j$ ,  $F_{ij}$  is the friction factor for travelling between zone  $i$  and  $j$ .  $F_{ij}$  is defined below:

$$F_{ij} = \alpha e^{-\beta d_{ij}} \quad (2)$$

where  $d_{ij}$  is the Haversine distance between the centroid of zone  $i$  and zone  $j$ .  $\alpha$  is set to 1. And  $\beta$  is calibrated with the OD matrix directly derived from the raw data so that they are optimally approximated by the estimated OD matrix where the similarity is measured by SpSSIM. Trip distribution uses Iterative Proportional Fitting (IPF) to assign trips from the predefined productions and attractions to the estimated OD matrix (Ben-Akiva et al. 1985; McCord et al. 2010). All the OD matrices are standardised, so that every cell has a value between 0 and 1 representing the probability of the connection between two zones.

### Evaluation of Twitter OD matrices

The comparison techniques used to evaluate the Twitter OD matrices include the visualisation of the OD matrices and the similarity measure (SpSSIM) between the OD matrix from Twitter and from the external sources (“[Spatially weighted structural similarity index](#)”

section). An essential aspect of human mobility behaviour is the travel distance ( $d$ , km) of OD pairs, whose distribution provides another facet for validating Twitter to estimate travel demand. Therefore, we compare this distribution from Twitter data with that from other sources whose similarity is measured by Kullback-Leibler (KL) divergence measure (Wang et al. 2019; Smolak et al. 2020). The smaller the KL divergence, the more similar are the two given distributions.

### Spatially weighted structural similarity index

To evaluate the feasibility of using Twitter data as a proxy for travel demand, we compare Twitter-based OD matrices with those from external data sources and measure their similarity. The more similar, the better Twitter data work as a source for travel demand estimation.

Originally proposed by Wang et al. (2004), the structural similarity (SSIM) measures the similarity between two images for assessing image quality. We use the spatially weighted structural similarity index (SpSSIM) (Jin et al. 2019) that was later introduced to the field of transport for comparing the quality of OD matrices that are based on different data sources (Djukic et al. 2013; Pollard et al. 2013). This newly proposed SpSSIM overcomes the SSIM's sensitivity issue due to the ordering of OD pairs, as raised by earlier studies (e.g., Djukic 2014).

The distances between all possible OD pairs create a matrix  $\mathbf{D}$  where  $d_{ij}$  is the Haversine distance between the centroids of zone  $i$  and zone  $j$ . The distances are binned by their percentile, yielding multiple groups of the same number of OD pairs (10% per group) based on their spatial adjacency ( $[D_{min}^b, D_{max}^b]$ ,  $b = 1, 2, \dots, n$ ). The spatial filtering matrix ( $\mathbf{W}^b$ ) consists of 0 and 1, and is defined as:

$$W_{ij}^{[D_{min}^b, D_{max}^b]} = \begin{cases} 1 & D_{min}^b \leq d_{ij} < D_{max}^b \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Two OD matrices,  $\mathbf{X}$  and  $\mathbf{Y}$ , are in a probabilistic form i.e., each cell is a value between 0 and 1 indicating the strength/probability of the connection between zone  $i$  and zone  $j$ . For a distance group, all the cells beyond the range of the distance are set to zero, with  $\mathbf{W}^b \mathbf{X}$  and  $\mathbf{W}^b \mathbf{Y}$  indicating the Hadamard product between  $\mathbf{W}^b$  and  $\mathbf{X}$  and  $\mathbf{W}^b$  and  $\mathbf{Y}$ , respectively. The similarity between  $\mathbf{X}$  and  $\mathbf{Y}$  on this distance range is indicated by  $\text{SpSSIM}(\mathbf{X}, \mathbf{Y}, \mathbf{W}^b)$  which is:

$$\frac{(2\mu_{\mathbf{W}^b \mathbf{X}} \mu_{\mathbf{W}^b \mathbf{Y}} + C_1)(2\sigma_{\mathbf{W}^b \mathbf{X}, \mathbf{W}^b \mathbf{Y}} + C_2)}{(\mu_{\mathbf{W}^b \mathbf{X}}^2 + \mu_{\mathbf{W}^b \mathbf{Y}}^2 + C_1)(\sigma_{\mathbf{W}^b \mathbf{X}}^2 + \sigma_{\mathbf{W}^b \mathbf{Y}}^2 + C_2)} \quad (4)$$

where  $\mu$  is the mean,  $\sigma$  the variance or the covariance between two matrices, and  $C_1$  and  $C_2$  are two constants.

The share of the travel demand for a distance group  $b$  is expressed below:

$$s^b = \sum_{i=1}^N \sum_{j=1}^N \mathbf{W}^b(i, j) \mathbf{X}(i, j) \quad (5)$$

where  $N$  indicates the number of traffic zones and  $s^b$  has a value between 0 and 1 indicating the share of trips that are expected to happen within the distance range  $b$ . The OD

pairs of different distance groups have imbalanced share of travel demand between them ( $s^b, b = 1, 2, \dots, n$ ). Accounting for the share of travel demand, the similarity between the two OD matrices ( $\mathbf{X}$  and  $\mathbf{Y}$ ) is quantified by SpSSIM as calculated below aggregating over all the distance groups:

$$\text{SpSSIM}(\mathbf{X}, \mathbf{Y}) = \sum_{b=1}^n \text{SpSSIM}(\mathbf{X}, \mathbf{Y}, \mathbf{W}^b) \cdot s^b \quad (6)$$

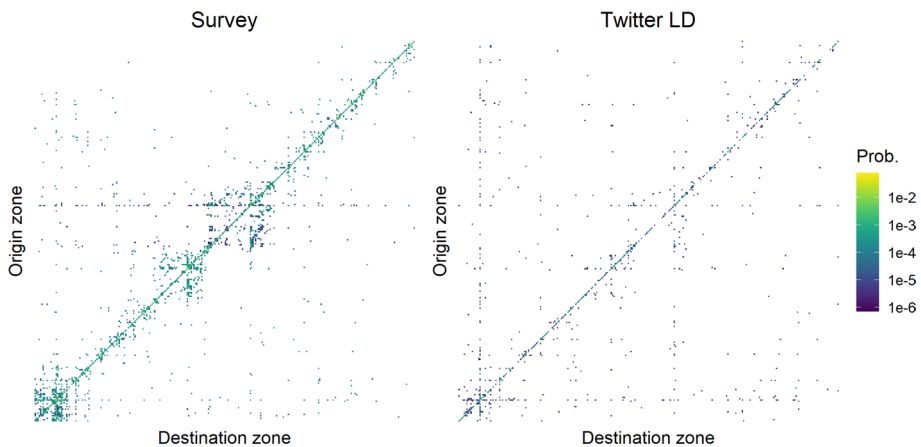
For the matrices in this study, have a mean squared  $\mu^2 \sim 10^{-12}$ – $10^{-9}$  and  $\sigma \sim 10^{-9}$ – $10^{-6}$ , hence we adjust the constants to values of  $C_1 = 10^{-16}$ – $10^{-13}$ ,  $C_2 = 10^{-11}$ – $10^{-8}$ . The detailed justification of these selections can be found in a study by Pollard et al. (2013). From the above definition, SpSSIM has a value between 0 and 1. SpSSIM equals to 1 when two OD matrices have the same exact pattern.

## Results

### Trip records comparison

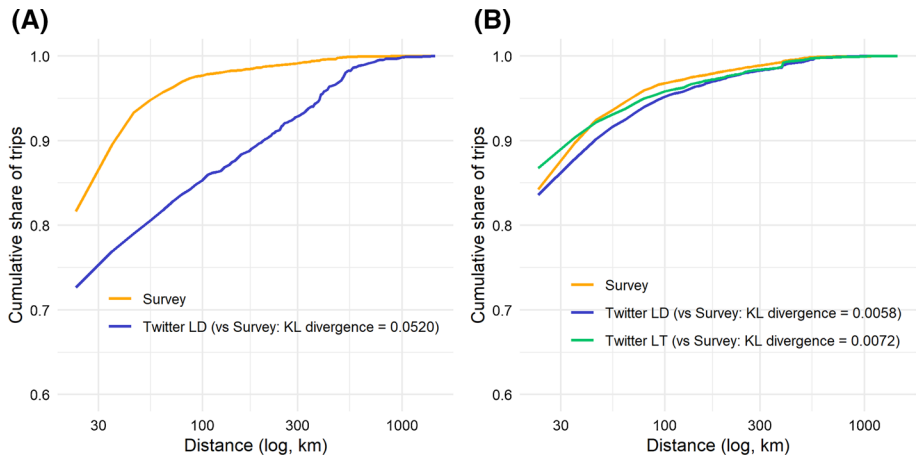
#### Commuting trips

The commuting OD matrices of Survey and Twitter LD are shown in Fig. 4. The diagonal cells have higher values indicating that most individuals commute within their residence municipality. Spatial proximity between municipalities (neighbouring cells) also affects the inter-municipality commuting flows in Survey's OD matrix. However, this is not the same for the OD matrix derived from Twitter LD: the estimated home and workplace are not strongly influenced by the distance between them. The similarity between Twitter LD and Survey is low (SpSSIM = 0.39, KL divergence = 0.052).



**Fig. 4** Commuting origin (y-axis) and destination (x-axis) OD matrices based on Survey (left figure) and Twitter LD (right figure) trip records. Colour showing the probability of connections between each pair of zones, i.e., the proportion of trips. (Color figure online)





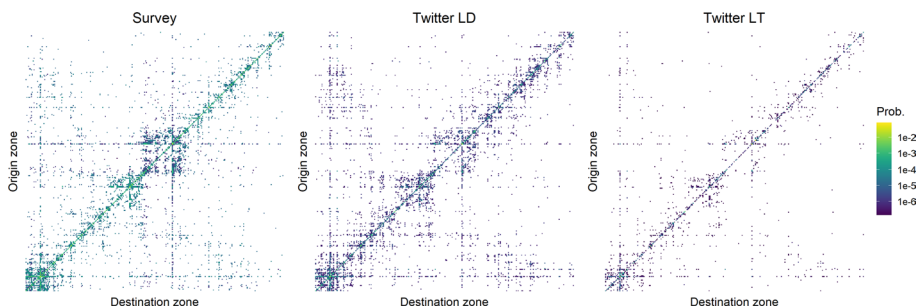
**Fig. 5** Travel distance distribution of data for Twitter and Survey. Cumulative share of trips is the probability of travelling between zones at a distance equal or below a certain threshold. **a** Commuting trips. **b** Weekday trips. (Color figure online)

Commuting travel distances are shown in Fig. 5a. The distribution produced by Twitter LD is significantly higher than Survey. This is consistent with the observation in Fig. 4 that Twitter LD does not capture the frequent commuting flows between neighbouring cells as shown in Survey’s OD matrix.

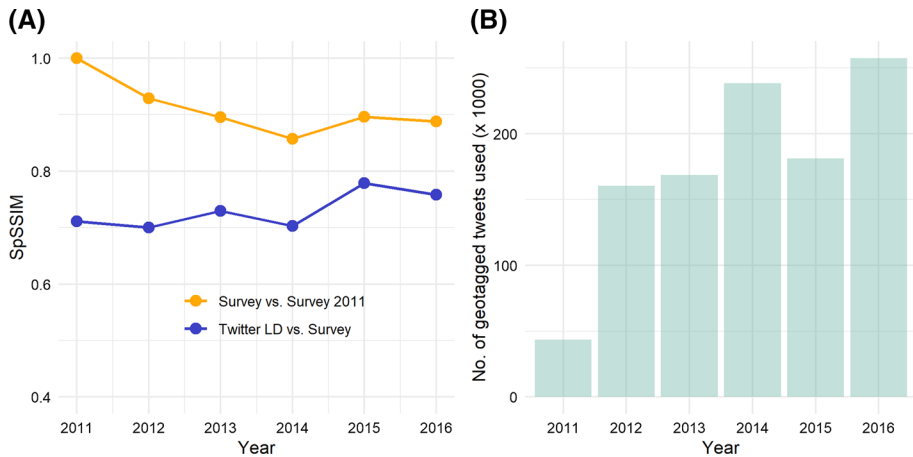
## Weekday trips

The weekday trips’ OD matrices (Fig. 6) from the two Twitter datasets and Survey are visually more similar than the results of commuting travel demand shown in “[Commuting trips](#)” section. Twitter data produce sparse matrices, and Twitter LT is sparser than Twitter LD. In addition, Twitter LD looks more similar to Survey as compared to Twitter LT.

The quantitative similarity (SpSSIM) results are in line with the visual results; at the national level, the trips converted from displacements in Twitter LD approximate the Survey OD matrix better than those from Twitter LT (0.87 vs. 0.64). Still at the national level, the Twitter trips collected with either sampling method approximate the Survey’s trip



**Fig. 6** OD matrices based on the weekday trip records from Survey and Twitter data



**Fig. 7** Similarity between Survey and Twitter LD and its sample size by year. **a** Survey vs. Twitter LD over time. The curve of Survey shows how the OD matrix deviates from the baseline year, 2011 for Survey records. **b** Number of geotagged tweets in Twitter LD over time

**Table 2** The similarity between the modelled OD matrices using Twitter data and Samplers traffic model's outputs

Scale	Model	Twitter	SpSSIM	KL divergence
City	A	LD	0.74	0.072
		LT	0.54	0.219
	B	LD	0.80	0.023
		LT	0.85	0.017
National	A	LD	0.52	0.317
		LT	0.40	0.364
	B	LD	0.54	0.021
		LT	0.54	0.026

Model A: displacement conversion plus gravity model; Model B: density-based approach plus gravity model. For all models,  $\beta = 0.03$

distance distribution well (Fig. 5b) in contrast with a greater discrepancy observed in commuting travel distances (Fig. 5a).

The similarity of trips disaggregated by year is rather stable, when comparing Survey to the baseline year (2011), and Twitter LD to the Survey over time (see Fig. 7). The stability of this similarity between Twitter LD and Survey suggests that Twitter data are reliable in capturing changes in travel and thus are suitable for estimating the change in national-level travel demand over time.

## Model outcomes

When looking at the similarity with Samplers, Twitter data generally work well at the city level (0.54 to 0.85), while the performance at the national level is not as good (0.40 to 0.54), see Table 2. The sampling method matters; Twitter LD is more similar to Samplers than Twitter LT, especially when using Model A with Displacement conversion (Twitter

LD<sup>A</sup> vs. Twitter LT<sup>A</sup>, National = 0.52 vs. 0.40, and City = 0.74 vs. 0.54). Combining the density-based approach and the gravity model (Model B) produces better similarity results compared to using Displacement conversion (Model A) at both spatial scales. This is probably due to the fact that Model B manages to increase the number of available geotagged tweets five-fold relative to Model A.

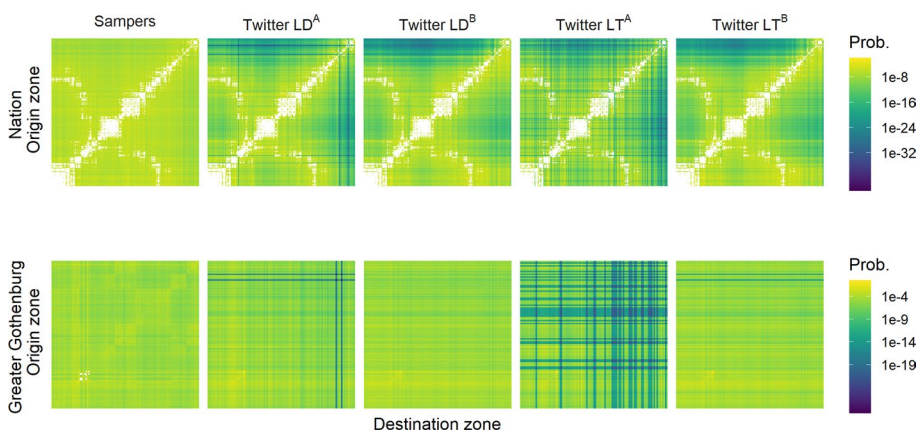
Figure 8 shows model outcomes. At both the national level (upper row) and city level (bottom row), the visualisation confirms the similarity as quantified by the SpSSIM and KL divergence values shown in Table 2. Comparing the two spatial scales, the greater number of traffic zones and larger geographical coverage make the national level more challenging to model using Twitter data due to the sparsity issue, leading to lower values of similarity in general than the city level.

Finally, we look into the travel distance distribution (Fig. 9). At the national level, the Twitter-based output with Model A shows greater short-distance travel demand when compared with the Sampers model (Fig. 9a, b). At the city level, however, the Twitter models approximate the traffic model better than at the national level in general, especially for Twitter LD (Fig. 9c, d). At both spatial scales, Model B represents the trip distance distribution better than Model A.

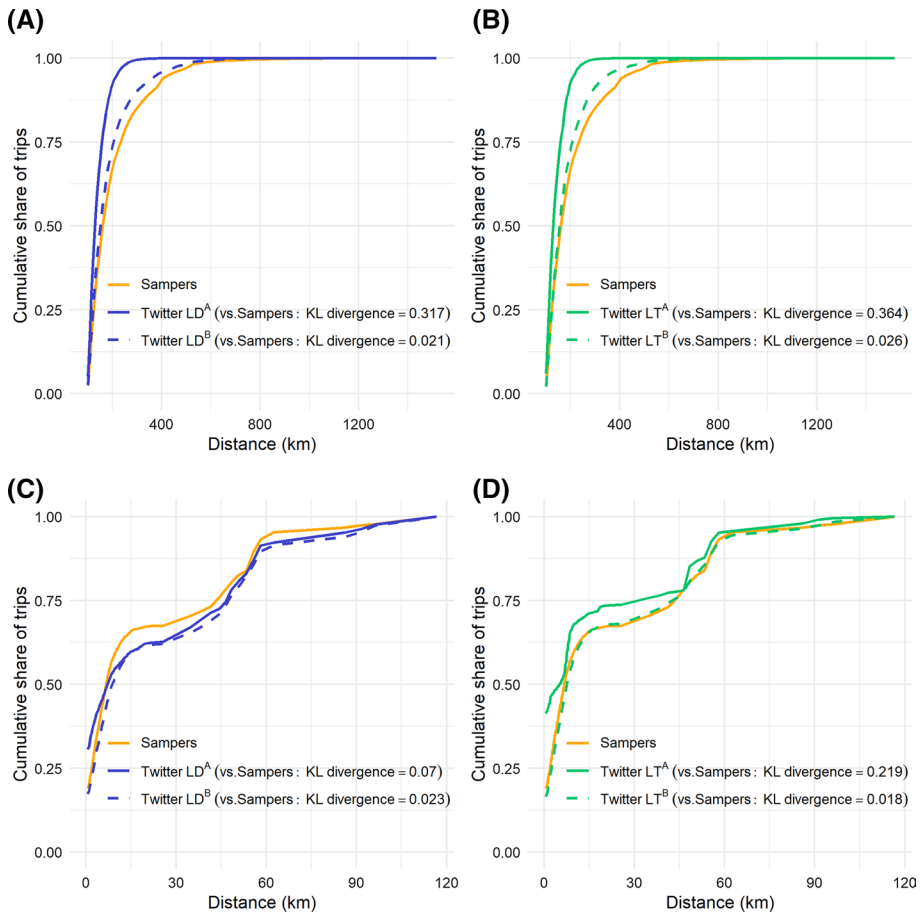
### Sensitivity of Twitter-based model outcomes to the sample size

Sensitivity of outcomes to sample size and to sampling method of tweets (LD or LT) are tested using a share of geotagged tweets from 1% to 99%, with a step length of 1% and 10 repetitions of random sampling, to create outputs using models A and B with the same settings as above. Figure 10 shows the similarity results.

As expected, as more geotagged tweets are included in the modelling, the similarity between the outputs of the Twitter-based OD matrix and the Sampers model increases and remains within a smaller range. The national level is more sensitive to data sparsity, because the number and the geographical coverage of traffic zones is greater than at the city level, therefore requires a greater number of tweets to reach a stable (but still lower) similarity. In terms of methodology, Model A is more sensitive to the number of geotagged



**Fig. 8** Estimated OD matrices by gravity model and Sampers' model outputs. A: Displacement conversion plus gravity model. B: Density-based approach plus gravity model

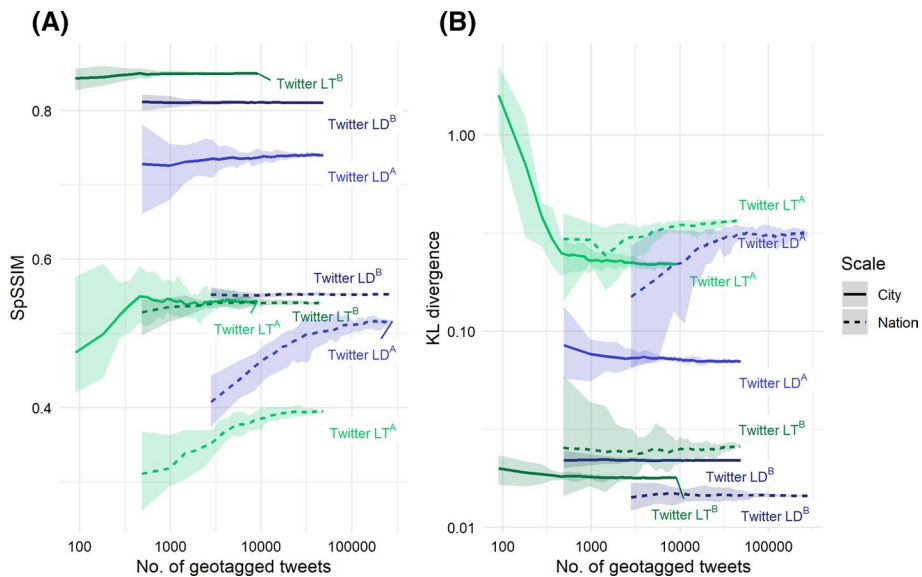


**Fig. 9** Trip distance distribution. Cumulative share of trips refers to the probability of travel between zones below a given distance. The trip distance is from the estimated OD matrices by *A* displacement conversion plus gravity model and by *B* density-based approach plus gravity model. **a** National level—Twitter LD. **b** National level—Twitter LT. **c** City level—Twitter LD. **d** City level—Twitter LT

tweets than Model B, especially with respect to the stability of the results with a smaller number of tweets, and is generally associated with poorer results.

## Discussion

Based on the comparison with travel surveys and the government's traffic simulation model, our study suggests that geotagged tweets can be suitable for estimating the overall travel demand (including OD matrix and travel distance) for an average weekday. However, as discussed in “[Commuting travel demand estimation](#)” section, estimation of the commuting travel demand is not reliable even though the data have been used for this purpose in the literature (e.g., Zagatti et al. 2018). We further discuss the impact of spatial scale (“[The impact of spatial scale](#)” section), sampling size and sampling method (“[The impact of sampling size and methods of data collection](#)” section). In “[A novel density-based approach:](#)



**Fig. 10** Similarity, **a** SpSSIM and **b** KL divergence, as a function of the number of geotagged tweets. Green colours show the results using Twitter LT and blue colours show the results using Twitter LD. For the 10 model runs of each tweets sample size, the curve shows the average value of SpSSIM/KL divergence and the shaded area shows the maximum and minimum value of SpSSIM/KL divergence. Model A—displacement conversion plus gravity model; Model B—density-based approach plus gravity model. For all models,  $\beta = 0.03$ . (Color figure online)

geotagged tweets as attractions generators as opposed to trips generators” section, we discuss the clear advantage of the innovative density-based approach proposed here and offer possible explanations for this result.

## Commuting travel demand estimation

The reliability of estimated commuting trips using geotagged tweets is low.

We use a simple, yet commonly adopted method (Wang et al. 2018) to identify work-place locations: the most visited non-home location during 7 am–8 pm on weekdays. However, not all Twitter users are employed (according to the OECD, employment rate is 77.1% in Sweden for those aged 15–64) and not all work outside of home between 8 am–8 pm. Despite the difference between Twitter users and the general population, one can expect that the aggregate commuting trips should be quite similar given our findings with regards to overall trips. The observed dissimilarity could be explained by the errors introduced using the simple method mentioned above, and the behaviour biases of Twitter users. Most Twitter users may not feel comfortable or interested in geolocating their homes and workplaces online publicly due to privacy concerns. A geotag usage survey based on 400 US residents shows that 70% of their geotags happen in places that people visit infrequently (Tasse et al. 2017). The temporal distribution of geotagging behaviour resembles that of a leisure activity pattern (Federal Office for Spatial Development ARE 2017). Moreover, geotag users tend to geotag locations that are not within their neighbourhood; and the geotagged locations concentrate substantially at locations farther away from

the daily mobility area (Tasse et al. 2017). This suggests Twitter can have significant shortcomings when used for capturing routine activities such as trips between home and the workplace.

## The impact of spatial scale

The main obstacle of using Twitter data at a large spatial scale is the sparsity.

Using geotagged tweets for travel demand estimation requires a sufficient sample size, which depends on the form of Twitter data (“[The impact of sampling size and methods of data collection](#)” section), penetration rate, and the number of samples collected. An important contribution of our work is to examine the robustness of the travel demand estimation for different spatial scales. Most studies have focused on one particular scale, be it at the city (Wang et al. 2018) or international (Hawelka et al. 2014) level. Yet, the validity of the selection of scale for the purpose of travel demand estimation remains unclear until different spatial scales are properly investigated. Despite the accumulation of geotagged tweets over months (Twitter LT) to years (Twitter LD), the share of zones with insufficient coverage increases at the national level, due to Twitter’s lower penetration rate outside urban centres. Therefore, using geotagged tweets for travel demand estimation requires appropriate selection of spatial aggregation i.e., zoning system.

## The impact of sampling size and methods of data collection

The more geotagged tweets included in the modelling, the better Twitter is at estimating travel demand. Twitter LD results in a much larger number of geotagged tweets that overall better represents population mobility patterns.

The number of geotagged tweets in Twitter LT is 30% of those in Twitter LD (3.4 vs. 11.5 million geotagged tweets). Compared with Twitter LT, Twitter LD covers longer period (9 years compared with 6 months for Twitter LT) with fewer users (2311 compared with 24,442 with Twitter LT). Our study demonstrates that, however, the long-term coverage of longitudinal geotagged tweets by top users (User Timeline API) compensates for the time sparsity and helps to recreate a more complete picture of population mobility patterns, and therefore, is more reliable for travel demand estimation than the lateral dataset (Twitter LT). However, this gap narrows or disappears when using a novel density-based approach developed in this study, see below in “[A novel density-based approach: geotagged tweets as attractions generators as opposed to trips generators](#)” section.

## A novel density-based approach: geotagged tweets as attractions generators as opposed to trips generators

The density-based approach utilises more geotagged tweets, resulting in better representation of travel demand.

The common practice of adding a time threshold (Model A, displacement method) to capture “trips” drastically reduces the available Twitter data for travel demand estimation: only 20–35% of geotagged tweets are utilised to estimate the overall travel demand. This reduction limits the application of geotagged tweets given that sparsity is already one of its key drawbacks.

Without the need for a time threshold, the density-based approach (Model B) increases usable data by 2–7 times. This drastically increases the similarity scores of the OD matrices compared with Sampers' model outputs and the method is not so sensitive to sample size. Considering both similarity and stability in Fig. 10, a magnitude of 1000 geotagged tweets is sufficient for the city-level and the national level requires 10,000 tweets to reach a stable similarity. This is equivalent to a minimum of 1 geotagged tweet per 1000 persons for the entire Sweden and 2 geotagged tweets per 1000 persons for the more densely populated city, requiring roughly an order of magnitude fewer samples than the other method.

Not only does the density-based approach produce better OD matrices, it also produces better trip distance distributions compared with the Surveys. Tasse et al. (2017) suggest that most Twitter users geotag their tweets within an hour of arrival (if at all), thus geotagging may be a timely indicator of the start time of the activity. This emphasises that Twitter users geotweet to report activities instead of trips, therefore, the density of geotagged tweets naturally reflects the attractiveness of zones. This motivates the proposed density-based approach, which regards the tweets density of zones as the attractions and the population size of zones as the productions. The results suggest that the density-based approach captures some of the population flows, because we assume that the generated trips between zones are proportional to (1) the population and (2) the number of activities some of which are geotagged.

A plausible explanation could be that the improvement was solely ascribed to the use of population as production, instead of the use of geotagged tweets as attractions. To test this assumption, we observe the changes in similarity metric when we assign attraction and production using: (1) population count as both production and attraction and (2) geotagged tweets count as both production and attraction. They both perform better than the displacement conversion; however, they are not as good as the density-based approach.

The density-based approach can be extended to compute time-dependent attractions by aggregating geotagged tweets across different temporal profiles, providing a dynamic picture of travel demand by time of day, week, or season.

## Conclusion

This study critically examines the feasibility of using geolocations of Twitter data to estimate population mobility. The overall results suggest that Twitter data can be suitable for modelling the overall travel demand for an average weekday but not the commuting travel demand due to the low reliability of identifying home and work locations. This makes it hard to replace the conventional national travel surveys and similar survey methods, in which users report the purposes of all trips.

The key strengths of social media data are that they are low-cost, abundant, available in real-time, and free of arbitrary geographical partition. However, there are also significant limitations: population and behavioural biases and lack of important information such as social demographic information and trip purposes. Despite clear indications of overrepresenting residents in big cities and their leisure activities from the existing literature, we demonstrate in the present study that geotagged tweets can provide a reasonably good travel demand estimation that also captures the trends over time.



## Limitations and outlook

The present study uses data from a single country. Further exploration is needed to understand how the findings can be generalised to the other regions, despite comparable validation data being difficult to come by. The proposed density-based approach allows flexible temporal aggregation to estimate time-varying travel demand, however, our validation data only represent an average day without any time-dependent demand estimation and exclude weekends and long-distance travel. This limits our ability to test the validity of time-dependent travel demand estimates and travel outside of the validation regions. One future direction is to test the performance of the density-based approach at different levels of temporal and spatial resolution. In addition, future work can use the tweet contents, not used in the current study, as this can provide additional information for inferring trip purposes. Last but not least, despite adjusting the Twitter users in Twitter LD based on its ratio to the true population at the municipality level, the method is simple and can be further improved, such as considering additional socio-demographic dimensions, to better represent the population. In the future, we can explore methods such as inferring trip purposes and socio-demographic information to derive more robust and more reliable travel demand estimation.

**Authors' contributions** Yuan Liao, Sonia Yeh and Jorge Gil conceptualised the study. All the authors designed the methods. Yuan Liao analysed the data. All the authors wrote the manuscript.

**Funding** Open Access funding provided by Chalmers University of Technology. The authors acknowledge the financial support of the Swedish Research Council for Sustainable Development (Formas, project number 2016-01326).

## Compliance with ethical standards

**Conflicts of interest** The authors declare no competing financial or non-financial interests.

**Availability of data and material** The aggregate data that support the findings of this study are available upon request from the corresponding author. The individual data can not be made available since they contain information that could compromise the privacy of individuals.

**Code availability** Not applicable.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abbasi, A., Rashidi, T.H., Maghrebi, M., Waller, S.T.: Utilising location based social media in travel survey methods: bringing twitter data into the play. In: Proceedings of the 8th ACM SIGSPATIAL international workshop on location-based social networks, pp. 1–9 (2015)
- Anda, C., Erath, A., Fourie, P.J.: Transport modelling in the age of big data. *Int. J. Urban Sci.* **21**(sup1), 19–42 (2017)
- Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M.: Human mobility: models and applications. *Phys. Rep.* **734**, 1–74 (2018)
- Ben-Akiva, M., Macke, P.P., Hsu, P.S.: Alternative methods to estimate route-level trip tables and expand on-board surveys. *Transp. Res. Record* **1037**, 1 (1985)
- Calabrese, F., DiLorenzo, S., Liu, L., Ratti, C.: Estimating origin–destination flows using opportunistically collected mobile phone location data from one million users in Boston metropolitan area. *IEEE Pervasive Comput.* **10**(4), 36–44 (2011)
- Cheng, Z., Jian, S., Rashidi, T.H., Maghrebi, M., Waller, S.T.: Integrating household travel survey and social media data to improve the quality of od matrix: a comparative case study. *IEEE Trans. Intell. Transp. Syst.* **21**(6), 2628–2636 (2020)
- Djukic, T.: Dynamic OD demand estimation and prediction for dynamic traffic management. TRAIL Thesis Series no. T2014/9, Delft University of Technology, 2600 GA Delft, The Netherlands (2014)
- Djukic, T., Hoogendoorn, S., Van Lint, H.: Reliability assessment of dynamic od estimation methods based on structural similarity index. Technical Report (2013)
- Ek, K., Wennerberg, E.: Estimating travel demand from twitter using an individual mobility model: in Sweden, the Netherlands and São paulo. Master's thesis, Chalmers tekniska högskola, S-412 96 Gothenburg, Sweden, <https://hdl.handle.net/20.500.12380/301742> (2020)
- Federal Office for Spatial Development ARE: Population's transport behaviour 2015. Technical Report (2017)
- Gao, S., Yang, J.A., Yan, B., Hu, Y., Janowicz, K., McKenzie, G.: Detecting origin–destination mobility flows from geotagged tweets in greater Los Angeles area. In: Eighth International Conference on Geographic Information Science (GIScience'14), Citeseer (2014)
- Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. *Nature* **453**(7196), 779–782 (2008)
- Hasan, S., Ukkusuri, S.V.: Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C: Emerg. Technol.* **44**, 363–381 (2014)
- Hasnat, M.M., Hasan, S.: Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transp. Res. Part C: Emerg. Technol.* **96**, 38–54 (2018)
- Hasnat, M.M., Faghih-Imani, A., Eluru, N., Hasan, S., et al.: Destination choice modeling using location-based social media data. *J. Choice Modell.* **31**, 22–34 (2019)
- Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C.: Geo-located twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **41**(3), 260–271 (2014)
- Jeuken, G.S.: Using big data for human mobility patterns—examining how twitter data can be used in the study of human movement across space. Master's thesis (2017)
- Jin, P., Cebelak, M., Yang, F., Zhang, J., Walton, C., Ran, B.: Location-based social networking data: exploration into use of doubly constrained gravity model for origin–destination estimation. *Transp. Res. Record J. Transp. Res. Board* **2430**, 72–82 (2014)
- Jin, C., Nara, A., Yang, J.A., Tsou, M.H.: Similarity measurement on human mobility data with spatially weighted structural similarity index (SpSSIM). *Trans. GIS* **24**(1), 104–22 (2020)
- Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D.: Understanding human mobility from twitter. *PloS One* **10**(7), e0131469 (2015)
- Kheiri, A., Karimipour, F., Forghani, M.: Intra-urban movement flow estimation using location based social networking data. *Int. Arch. Photogr. Remote Sens. Spatial Inf. Sci.* **40**(1), 785 (2015)
- Lee, J.H., Gao, S., Goulias, K.G.: Can twitter data be used to validate travel demand models. In: 14th International Conference on Travel Behaviour Research (2015)
- Lee, J.H., Davis, A., McBride, E., Goulias, K.G.: Statewide comparison of origin–destination matrices between California travel model and twitter. *Mobility Patterns, Big Data and Transport Analytics*, pp. 201–228. Elsevier, Amsterdam (2019)
- Lenormand, M., Picornell, M., Cantú-Ros, O.G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frias-Martinez, E., Ramasco, J.J.: Cross-checking different sources of mobility information. *PLoS One* **9**(8), e105184 (2014)

- Lenormand, M., Gonçalves, B., Tugores, A., Ramasco, J.J.: Human diffusion and city influence. *J. R. Soc. Interface* **12**(109), 20150473 (2015)
- Liao, Y., Yeh, S.: Predictability in human mobility based on geographical-boundary-free and long-time social media data. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 2068–2073 (2018)
- Liao, Y., Yeh, S., Jeuken, G.S.: From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data. *EPJ Data Sci.* **8**(1), 34 (2019)
- Maghrebi, M., Abbasi, A., Rashidi, T.H., Waller, S.T.: Complementing travel diary surveys with twitter data: application of text mining techniques on activity location, type and time. In: 2015 IEEE 18th international conference on intelligent transportation systems (ITSC). IEEE, pp 208–213 (2015)
- McCord, M.R., Mishalani, R.G., Goel, P., Strohl, B.: Iterative proportional fitting procedure to determine bus route passenger origin–destination flows. *Transp. Res. Record* **2145**(1), 59–65 (2010)
- McNally, M.G.: The four-step model. In: Hensher, D.A., Button, K.J. (eds.) *Handbook of Transport Modelling*, vol. 1, pp. 35–53. Emerald Group Publishing Limited (2007). <https://doi.org/10.1108/9780857245670-003>
- Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N.: Understanding the demographics of twitter users. *ICWSM* **11**(5), 25 (2011)
- Molloy, J., Moeckel, R.: Improving destination choice modeling using location-based big data. *ISPRS Int. J. Geo-Inf.* **6**(9), 291 (2017)
- Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? Comparing data from twitter’s streaming API with twitter’s firehose. In: *ICWSM* (2013)
- Official Statistics of Sweden (2016) Swedish National Travel survey (RVU Sweden) 2011–2016. <https://www.trafa.se/en/travel-survey/travel-survey/>
- Osorio-Arjona, J., García-Palomares, J.C.: Social media and urban mobility: using twitter to calculate home-work travel matrices. *Cities* **89**, 268–280 (2019)
- Pollard, T., Taylor, N., van Vuren, T., MacDonald, M.: Comparing the quality of OD matrices in time and between data sources. In: *Proceedings of the European Transport Conference* (2013)
- Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S.: Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. *Transp. Res. Part C: Emerg. Technol.* **75**, 197–211 (2017)
- Ribeiro, A.I.J.T., Silva, T.H., Duarte-Figueiredo, F., Loureiro, A.A.: Studying traffic conditions by analyzing foursquare and instagram data. In: *Proceedings of the 11th ACM symposium on Performance evaluation of wireless ad hoc, sensor, and ubiquitous network*. ACM, pp. 17–24 (2014)
- Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C.: Unravelling daily human mobility motifs. *J. R. Soc. Interface* **10**(84), 20130246 (2013)
- Smolak, K., Rohm, W., Knop, K., Sıla-Nowicka, K.: Population mobility modelling for mobility data simulation. *Comput. Environ. Urban Syst.* **84**, 101526 (2020)
- Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. *Science* **327**(5968), 1018–1021 (2010)
- Tasse, D., Liu, Z., Sciuto, A., Hong, J.I.: State of the geotags: Motivations and recent changes. In: *ICWSM*, pp. 250–259 (2017)
- Twitter, Inc.: Filter realtime tweets. <https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter> (2019a)
- Twitter, Inc.: Get Tweet timelines. <https://developer.twitter.com/en/docs/tweets/timelines/overview> (2019b)
- Twitter, Inc.: Twitter provides Tweets and associated metadata including geo data, images, and mentions. <https://support.gnip.com/sources/twitter/> (2019c)
- Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
- Wang, Q., Phillips, N.E., Small, M.L., Sampson, R.J.: Urban mobility and neighborhood isolation in America’s 50 largest cities. *Proc. Natl. Acad. Sci.* **115**(30), 7735–7740 (2018)
- Wang, J.X., Huang, J., Duan, L., Xiao, H.: Prediction of Reynolds stresses in high-mach-number turbulent boundary layers using physics-informed machine learning. *Theor. Comput. Fluid Dyn.* **33**(1), 1–19 (2019)
- Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., Buckee, C.O.: The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface* **10**(81), 20120986 (2013)
- Yang, F., Jin, P.J., Cheng, Y., Zhang, J., Ran, B.: Origin–destination estimation for non-commuting trips using location-based social networking data. *Int. J. Sustain. Transp.* **9**(8), 551–564 (2015)
- Yue, Y., Lan, T., Yeh, A.G., Li, Q.Q.: Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. *Travel Behav. Soc.* **1**(2), 69–78 (2014)

- Zagatti, G.A., Gonzalez, M., Avner, P., Lozano-Gracia, N., Brooks, C.J., Albert, M., Gray, J., Antos, S.E., Burci, P., zu Erbach-Schoenberg, E., et al.: A trip to work: estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR. *Dev. Eng.* **3**, 133–165 (2018)
- Zhang, Z., He, Q., Zhu, S.: Potentials of using social media to infer the longitudinal travel behavior: a sequential model-based clustering method. *Transp. Res. Part C: Emerg. Technol.* **85**, 396–414 (2017)
- Zipf, G.K.: The  $p$  1  $p$  2/ $d$  hypothesis: on the intercity movement of persons. *Am. Soc. Rev.* **11**(6), 677–686 (1946)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.